THE UNIVERSITY of York

Discussion Papers in Economics

No. 1999/30

A Further Investigation of Selten's Measure of Predictive Success

by

John Hey

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

A Further Investigation of Selten's Measure of Predictive Success

John D. Hey*
Universities of Bari and York

December 7, 1999

Abstract

There are two basic ways of assessing the goodness of fit of theories to data - one based on stochastic theory (for example, the maximised likelihood in some form) and one based on deterministic theory, for example, Selten's Measure of Predictive Success. This paper explores the second of these and presents an application of Selten's Measure to the problem of comparing and ranking various theories of decision making under risk. The paper uses an experiment which was specifically designed to provide insight into the usefulness of this measure. Specifically the questions in the experiment were chosen to give a high degree of discrimination between the various theories being ranked by Selten's measure. However, in common with a previous application, it is found that the measure appears to fail because it has no mechanism for differentiating between observations inconsistent with the theories. This seems to be an inherent failing of a measure based on deterministic theory.

1 Introduction

When attempting to assess the relative 'goodness of fit' of theories competing to explain a given set of observations, there are two approaches that the scientist may employ: one based on some kind of stochastic story underlying the generation of the data, and one based on some kind of deterministic story. This paper concentrates on the latter approach, and, in particular, examines a measure of 'goodness of fit' proposed in this Review by Professor Selten (Selten, 1991). In this paper, Selten proposed a method for comparing the predictive success of theories which differ in terms of their parsimony and predictive power.

*I am grateful to the Economic and Social Research Council of the UK for a grant (R000 23 6636) "Experimental Investigations of Errors in Decision Making" which financed the experiments reported in this paper. I am also grateful to Marie-Edith Bissey and Vittoria Levati, Research Fellows on this grant, who assisted me in the running of the experiments. I also benefitted from discussions with Marie Bissey concerning the analysis of the data.

An application of this measure was presented in (Hey, 1998). There it was found there that Selten's Measure did not appear to be a very useful tool for discriminating between theories. However, it could be argued that the experiment on which it was based was not designed in a way that fully demonstrated the virtues of this measure for comparing the 'goodness of fit' of competing theories. This current paper attempts to remedy this deficiency by basing the analysis on an experiment specifically designed to allow discrimination between theories.

The paper proceeds as follows: in the next section Selten's Measure of Predictive Success is described in general terms and then in terms of the particular field of application studied in this paper. The following two sections then describe the design and implementation of the experiment. Then the Measure is applied to the data generated by the experiment. Finally some concluding remarks are made.

2 Selten's Measure of Predictive Success

Selten's measure trades off the *predictive parsimony* of a theory against the *descriptive power* of a theory. The *descriptive power* of a theory is measured by the variable r which is the proportion of actual observations consistent with the theory; the *predictive parsimony* is measured by the variable a which is the proportion of all possible outcomes that are consistent with the theory. Clearly the higher is r the better and the lower is a the better. Any measure of predictive success should therefore be an increasing function of r and a decreasing function of a. Selten provides (very reasonable) axioms which guarantee that the appropriate measure is given by

$$s = r - a \tag{1}$$

So, for example, if there are 100 possible outcomes in total but just 5 predicted by that theory, then $a = 0.05$; further if there are 1000 observations and of those 1000, 680 are in the set of 5 outcomes predicted by that theory (with the remaining 320 in the set of 95 outcomes not predicted by the theory), then $r = 0.68$. Thus $s = r - a = 0.68 - 0.05 = 0.63$. Selten's suggested procedure is to calculate s for all competing theories and declare as the

best that theory for which s is highest.

Here this measure is used to compare various competing theories of decision making under risk. The set of such theories is now very large (see (Camerer, 1995)) for an overview and evaluation) so here attention is restricted to a subset of these theories. Experimental data is used to shed light on the relative descriptive validity of the various theories. Both pairwise choice and complete ranking data are used (in common with a previous study by (Hey, 1998)) but in this instance the same subjects completed both the pairwise choice experiment and the complete ranking experiment

Consider first pairwise choice data and, in particular, an experiment in which subjects are asked n such pairwise choice questions. Suppose subjects are asked to express a preference and are not allowed to express indifference. Then on each of the n questions there are 2 possible responses and hence altogether there are 2^n possible responses over the experiment as a whole. Which of these 2^n outcomes is consistent with a particular theory depends upon the theory and, of course, on the set of questions being asked of the subjects. Usually it is the case (unless there is indifference) that risk-neutral behaviour will predict one and only one possible response - since there is only one choice on each pairwise choice which gives the highest expected return. A less restrictive theory, such as Expected Utility theory, usually permits more outcomes - depending on the questions. Even less restrictive theories, such as the various generalisations and extensions of Expected Utility theory, permit yet more outcomes - once again, depending on the particular choice of the theories.

Consider now a complete ranking experiment, and, in particular, consider an experiment in which the subject is asked to rank m risky choices in order of preference. Then there are $m!$ different possible orderings that the subject may express (though some of these may be ruled out if the subject is assumed to respect dominance). Again, risk-neutral behaviour is consistent with one and only one possible ordering (unless there is indifference). Again, less restrictive theories usually permit more orderings, the number depending upon the theory, and, of course, the risky choices.

Consider the general case in which the experiment is designed in such a way that the set of all possible responses is denoted by Ω. Suppose there are K theories competing to explain

the behaviour of subjects and index the various theories by k. Let S_k denote the set of possible responses consistent with theory k. Though the case of a theory which states that subjects are risk-neutral will not be considered in this paper, it is clear that for this theory, under the conditions stated above, the set S_k contains a single element. For a theory which states that a subject may do anything, the set S_k is equal to Ω. Let Expected Utility theory be $k = 1$. This is one of the competing theories examined in this paper. The other theories considered are all proper generalisations of Expected Utility theory in the sense that they reduce to Expected Utility theory under certain restrictions. It follows that $S_1 \subset S_k$ for all k. The set of outcomes that are consistent with both theory k_1 and theory k_2 is the set $S_{k_1k_2} = S_{k_1} \cap S_{k_2}$. It is clear that if $S_{k_1k_2} = S_{k_1} = S_{k_2}$ then the experiment is such that it is not possible to discriminate between the theories (on the basis, that is, of considering behaviour consistent with the two theories). If, on the contrary the set $S_{k_1k_2} = \emptyset$ then the experiment is potentially informative in terms of discriminating between the two theories (once again, on the basis of behaviour consistent with the two theories). Unfortunately, given that $S_1 \subset S_k$ for all k, it follows that $S_{k_1k_2} \neq \emptyset$. In this case, the best that can be hoped for is that $S_{k_1k_2} = S_1$. In general this will not be the case, but this suggests a criterion for choosing the set of risky choices in the experiment. More generally, the risky choices should be chosen in such a way that not only are all the $S_{k_1k_2}$ as small as possible, but also the higher order sets $S_{k_1k_2k_3} = S_{k_1} \cap S_{k_2} \cap S_{k_3}$ and so on. Clearly, once again, given that Expected Utility theory is nested within the other theories, it must be the case that $S_{k_1k_2k_3} \supset S_1$ and so on.

It should be clear from the above discussion that the choice of the questions affects the power of the experiment in discriminating between the various theories. It follows that the theories themselves should determine the choice of questions in the experiment. The next section describes the theories investigated in this experiment. This provides the necessary background for the following section, in which the choice of questions in the experiment is explained.

3 The theories Investigated in this experiment

In addition to Expected Utility theory, there are many theories of decision making under risk. Here a subset is investigated, with attention being restricted to those that appear, from previous work, to have the greatest empirical validity, and which also appear to have theoretical potential. Discussion here is minimal, being confined to a description of the functional forms; further detail and discussion can be found in (Hey, 1997). A two letter abreviation is used to identify the various preference functionals. The specifications listed below are those appropriate to this experiment, in which there were just 3 outcomes, denoted here by x_1, x_2 and x_3, with respective probabilities p_1, p_2 and p_3.

(1) Expected Utility theory[1] (**EU**)

$$V(\mathbf{p}) = p_2 u + p_3 \tag{2}$$

(2) Disappointment Aversion theory (**DA**)

$$V(\mathbf{p}) = \min(W_1, W_2) \tag{3}$$

where

$$W_1 = \frac{(1+\beta)p_2 u + p_3}{1 + \beta p_1 + \beta p_2} \tag{4}$$

and

$$W_2 = \frac{p_2 u + p_3}{1 + \beta p_1} \tag{5}$$

(3) Prospective Reference theory (**PR**)

$$V(\mathbf{p}) = \lambda(p_2 u + p_3) + (1-\lambda)(a_2 u + a_3) \tag{6}$$

where $a_i = |a_i|/(a_i n(\mathbf{p}))$ and $n(\mathbf{p})$ is the number of non-zero elements in $\mathbf{p}$.

[1]Note the normalisation: that the utility of the worst outcome is put equal to zero and that the utility of the best outcome is put equal to unity; the utility of the middle outcome is denoted by u.

(4) Rank Dependent Expected Utility theory - with Power weighting function (**RP**)

$$V(\mathbf{p}) = w(p_2 + p_3)u + w(p_3)(1 - u) \tag{7}$$

where $w(.)$ is the *power* function $w(p) = p^\gamma$.

(5) Rank Dependent Expected Utility theory - with Quiggin weighting function (**RQ**)

$$V(\mathbf{p}) = w(p_2 + p_3)u + w(p_3)(1 - u) \tag{8}$$

where $w(.)$ is the '*Quiggin*'[2] function $w(p) = p^\gamma/[p^\gamma + (1-p)^\gamma]^{(1/\gamma)}$.

(6) Weighted Utility theory (**WU**)

$$V(\mathbf{p}) = \frac{wp_2u + p_3}{p_1 + wp_2 + p_3} \tag{9}$$

It should be noted that **EU** involves just *one* parameter; u; while all the other theories involve *two* parameters: **DA**, u and β; **PR**, u and λ; **RP** and **RQ**, u and γ ; and **WU**, u and w. The number of parameters clearly affects the range and number of allowable responses by individuals having the respective preferences.

4 Designing the Experiment

The experiment was built on top of a basic set of m risky choices. All of these were risky choices involving three final outcomes, which are denoted by x_1, x_2 and x_3 where these are indexed in such a way[3] that $x_1 \prec x_2 \prec x_3$ where $\prec$ denotes 'less preferred than'. A specific risky prospect is now described by the three numbers p_1, p_2 and p_3 where p_i denotes the probability that the outcome will be x_i $(i = 1, 2, 3)$. Note, however, that these three numbers must sum to unity - which means that any risky prospect can be described by just two of these three numbers. Take p_1 and p_3 - respectively the probability of the worst outcome and the probability of the best outcome. Now employ the expositional

[2]But see also (Karmarkar, 1978) and (Karmarkar, 1979).

[3]We actually used amounts of money increasing in magnitude, so we are assuming that all our subjects preferred more money to less.

device known as the Marschak-Machina Triangle - with p_3 on the vertical axis and p_1 on the horizontal axis. See Figure 1. Each point within the Triangle represents some risky prospect; each of those on one of the sides of the Triangle is a prospect involving just two of the three outcomes; and those at the vertices of the Triangle are certainties (involving just one of the three outcomes). In the earlier experiment reported in (Hey, 1998) the probabilites were chosen to be multiples of 1/4. This automatically implied the choice of the 11 basic lotteries, labelled a through k in Figure 1. The remaining points on the 1/4th grid either dominate, or are dominated by, all the points chosen; for this reason they were excluded from the earlier experiment.

One conclusion from the earlier experiment was that it did not appear to be sufficiently discriminating. This was party a consequence of the fact that the risky choices used in the experiment - the points chosen in the Marschack-Machina Triangle - were not sufficienty dispersed around the Triangle. One way to see this is by considering the fact that the relative ranking by a subject of two of the risky choices depends upon the subject's local risk aversion in the region of the two choices. By this is meant[4] the slope of the individual's indifference curves in the Triangle between the two points representing the two risky choices.

Different theories imply different restrictions on the slopes of the indifference curves in the Marschak-Machina Triange. For example, Expected Utility theory restricts the slopes to be constant throughout the Triangle. In contrast, Weighted Utility theory allows the curves to fan out across the Triangle. Discrimination between the theories is thus achieved by designing an experiment that exploits these differences.

4.1 The Pairwise Choice part of the Experiment

In order to explain this, some detail needs to be provided. The actual questions in this experiment are used to provide the detail, but the discussion clearly can be generalised. Begin with the basic 11 risky choices illustrated in Figure 2. These, in fact, were the basic 11 risky choices used in this experiment. The reasons for their selection will be explained shortly, after a discussion of the implications of any particular selection has been given.

[4] See (Machina, 1982).

From these basic 11 risky choices, a set of pairwise choice questions can be constructed. Starting from 11 basic risky choices, it is clear that the number of possible pairwise choice questions is ${}_{11}C_2 = 55$. Some of these will involve a pair in which one of the two choices is *dominant* - as for example pair b and g in Figure 2. Such pairs were excluded from the Pairwise Choice part of this experiment - the reason being that typically such questions are uninformative: all the theories say that the dominant choice should be chosen and subjects' behaviour seems generally to be in accordance with this. Omitting such pairs leaves a total of 30 pairwise choice questions in which neither choice is dominant, for example c and k in Figure 2. On each of these pairwise choice questions, the individual's preferences will determine which of the two he or she prefers.

Consider, for example, Expected Utility theory. Consider a choice between two risky choices: $\mathbf{p} = (p_1, p_2, p_3)$ and $\mathbf{q} = (q_1, q_2, q_3)$. From Equation 2 above it is clear that the preference between $\mathbf{p}$ and $\mathbf{q}$ will be determined by the respective values of $p_2u + p_3$ and $q_2u + q_3$, where u is the individual's utility of the middle outcome. In particular, the individual will be indifferent between $\mathbf{p}$ and $\mathbf{q}$ if and only if $p_2u + p_3 = q_2u + q_3$, that is, if and only if:

$$u = -\frac{p_3 - q_3}{p_2 - q_2} \tag{10}$$

This determines a critical value for u. If the individual's value of u is equal to this critical value, then the subject is indifferent between $\mathbf{p}$ and $\mathbf{q}$; if the individual's value of u is greater that this critical value, then the subject prefers $\mathbf{p}$ to $\mathbf{q}$; if the individual's value of u is less than this critical value, then the subject prefers $\mathbf{q}$ to $\mathbf{q}$. Note that the value of the critical value given by Equation 10 depends upon the slope of the line joining the two risky choices under consideration, and, because pairs in which one choice is dominant are excluded from the experiment, is necessarily a number between 0 and 1.

For each pairwise choice there is such a critical value - depending upon the slope of the line joining the two points in the Marschak-Machina Triangle. With 30 pairwise choices in the Pairwise Choice part of the experiment, there are therefore potentially 30 such critical values. However, some of these coincide because the slopes of the lines joining the two points are the same for certain questions: for example, the critical value for the pair (e, h)

is 0.5, as is the critical value for the pair (f, k) - since the slopes of the respective lines are both equal (to 1). In fact, there are just 15 distinct critical values - corresponding to slopes of 1/4, 2/4 (=1/2), 3/4, 1/3, 2/3, 3/3, 4/3, 5/3, 2/2 (=1/1), 3/2, 4/2 (=2/1), 5/2, 6/2 (=3/1), 4/1, 5/1 and 6/1. The corresponding critical values are 0.2, 0.25, 0.333... (=1/3), 0.4, 0.428... (=3/7), 0.5, 0.571... (=4/7), 0.6, 0.625, 0.666... (=2/3), 0.714... (=5/7), 0.75, 0.8, 0.8333... (=5/6) and 0.857... (=6/7)[5].

These critical values are graphed (as vertical lines) in Figure 3. Along the horizontal axis are the permissable values - zero to unity - of the utitity parameter u, assuming that the permissable utility functions are restricted to those that are monotonically increasing[6]. Note that these 15 critical values divide up the space of permissable u values into 16 regions. To each pairwise choice question there corresponds a line in Figure 3; the individual's preferences on that question depend upon where the individual's u value lies in relationship to the corresponding line. It follows that the subject's preferences over the 30 pairwise choice questions in the Pairwise Choice part of the experiment depend upon in which of these 16 regions lies his or her u value. To each of these 16 regions corresponds a particular set of responses to the 30 pairwise choice questions. Conversely, it follows that for Expected Utility theory, there are only 16 permissable sets of responses to the Pairwise Choice part of the experiment.

Consider now an alternative theory, for example Weighted Utility theory. Again consider an individual whose preferences are in accordance with this theory choosing between two risky choices: $\mathbf{p} = (p_1, p_2, p_3)$ and $\mathbf{q} = (q_1, q_2, q_3)$. From Equation 9 above it is clear that the preference between $\mathbf{p}$ and $\mathbf{q}$ will be determined by the respective values of $\frac{wp_2u+p_3}{p_1+wp_2+p_3}$ and $\frac{wq_2u+q_3}{q_1+wq_2+q_3}$, where u is the individual's utility of the middle outcome and w is the individual's weight parameter (defined in Weighted Utility theory). In particular, the individual will be indifferent between $\mathbf{p}$ and $\mathbf{q}$ if and only if

$$\frac{wp_2u + p_3}{p_1 + wp_2 + p_3} = \frac{wq_2u + q_3}{q_1 + wq_2 + q_3} \tag{11}$$

[5]Note that the critical values are given by Equation 10 while the slopes are given by $(p_3 - q_3)/(p_1 - q_1)$.
[6]Recall the normalisation: $u(x_1) = 0, u(x_2) = u, u(x_3) = 1$.

This defines a curve in (w, u) space.

Consider now Figure 8. On the horizontal axis is graphed the utility parameter u, in the permissable range from 0 to 1. On the vertical axis is graphed the weight parameter w. From the theory itself, it is clear that w should be positive but there is no obvious upper bound. Here a rather arbitrary upper bound of 3 has been used. For each pairwise choice question, Equation 11 gives a curve in this space defining indifference between the two choices on that particular question: if the individual's u and w values are such that they lie on this line, then the individual is indifferent between the two choices on that pairwise choice question. Contrariwise, if the (u, w) pair lie off that line, then the individual has a strict preference (which depends upon which side of the line the individual falls). To each pairwise choice question there corresponds a curve in Figure 8; for an individual whose preferences are Weighted Utility then the individual's preferences on that question depend upon where the individual's (u, w) pair lies in relationship to the corresponding curve. Note that to each question in the Pairwise Choice part of the experiment there is a curve in Figure 8 though some may be coincident as was the case with Expected Utility theory. These 30 curves are graphed in Figure 8. It will be noted that these 30 curves divide up the permissable (u, w) space into a number of different regions. Given that the subject's preferences on any particular pairwise choice question depend upon where the individual's (u, w) value lies in relation to the curve for that question, it follows that the subject's responses to the 30 pairwise choice questions depend precisely upon in which of the various regions in the figure the individual's (u, w) value lies. To each region there corresponds a particular set of responses to the Pairwise Choice part of the experiment: different regions imply different responses. Conversely, any set of responses consistent with Weighted Utility theory must correspond to one of the regions in Figure 8.

A key question now is how many regions are there in Figure 8. Answering this question is crucial to the application of Selten's Measure of Predictive Success, as we need to know the number and hence the proportion of all possible outcomes that are consistent with each theory. Answering this question is not particular easy, either for Weighted Utility theory or for the other theories, as is obvious from Figure 8 and the other figures, which

give the corresponding graphs for the other theories: Figure 4 for Disappointment Aversion theory; Figure 5 for Prospective Reference theory; Figure 6 for Rank Dependent (with Power weighting function) theory; Figure 7 for Rank Dependent (with Quiggin weighting function) theory.

There are a number of ways to count the number of different regions in these figures. A crude way is to try and count them by eye - but this is difficult. An exact way is being developed by Marie Bissey (Bissey, 1997) but is not yet finalised. Here a crude grid search was used: the graphs on the figures were covered with a grid (of size dx in the horizontal direction and of size dy in the vertical direction) and at each grid point the preferences of the individual were calculated. Clearly the choice of dx and dy is crucial as some of the regions in some of the figures are particularly small. Given a choice of dx and dy a computer program calculated the number of distinctly different responses to the 30 pairwise choice questions and hence calculated the number of regions in the various spaces. The choice of dx and dy obviously affect both the computational time and the accuracy of the results (the number of regions found). In the results reported here, we used values of dx and dy at which no further increases in the number of regions were found. The number of regions found for the pairwise choice questions used in the experiment were as follows:

EU 16

DA 67

PR 161

RP 156

RQ 199

WU 150

Inevitably, Expected Utitity theory is the most restrictive of all the theories - allowing just 16 possible responses out of the grand total of $2^{30} = 1,073,741,824$ possible responses to the 30 pairwise choice questions. The other theories are less restrictive, though perhaps it is surprising how few are the extra responses permitted by these more general theories:

while the proportion of all possible outcomes that are consistent with the theory (Selten's a) is just 0.000000015 for Expected Utility theory, it rises to just 0.000000062 for **DA**, to 0.000000149 for **PR**, to 0.000000145 for **RP**, to 0.000000185 for **RQ** and to 0.000000149 for **WU**.

Clearly the choice of the questions in the experiment (both the number and their composition) influence the graphs illustrated in Figures 3 through 8, and in particular influence the numbers of permissable responses, as listed above. But there is a further factor to consider - the extent of the overlap between the responses consistent with the various theories and hence the discriminatory power of the experiment. For the risky choices used in the Pairwise Choice part of the experiment, the implications are shown in Table 1. In this table each row represents a particular combination of the theories (indicated by the asterisks in the final 6 columns) and the number in the first column indicates the number of responses (out of the total of $2^{30} = 1,073,741,824$ possible responses) consistent with that particular combination. For example, the first row shows that there are 16 possible responses consistent with all 6 theories. Naturally, given that **EU** is a special case of all the other theories, this is also the number of responses consistent with **EU** itself - as shown above.

An alternative way of summarising this information is to note that of the 429 responses consistent with one or more of the preference functionals, 253 are consistent with just *one* of the functionals, 106 are consistent with just *two*, 30 are consistent with just *three*, 22 are consistent with just *four*, 2 are consistent with *five* and 16 (the **EU** set) are consistent with all 6.

Returning now to the choice of the 11 basic risky choices used in the experiment, it should be noted that for any choice of 11 basic lotteries, there is a table corresponding to Table 1. Moreover the analysis in the paragraph above can be repeated for any set of 11 basic risky choices. Different sets of 11 basic risky choices give different degrees of overlap and different degrees of discrimination. How might one select the 'best' set of 11 basic risky choices?

The problem is that the Marschak-Machina Triangle of Figure 2 contains 45 possible

Table 1: Intersections between the permissable responses

number of permissable patterns	**EU**	**DA**	**PR**	**RP**	**RQ**	**WU**
16	*	*	*	*	*	*
2		*	*	*	*	*
13		*		*	*	*
1		*	*		*	*
8			*	*	*	*
2		*	*			*
8		*			*	*
1		*		*		*
7			*		*	*
2			*	*	*	
10				*	*	*
10		*				*
1		*	*			
48			*		*	
2			*			*
7				*		*
29				*	*	
9					*	*
13		*				
72			*			
68				*		
46					*	
54						*

candidates. These can be reduced to 41 if obviously dominating or dominated choices are omitted. But there are ${}_{41}C_{11} = 3,159,461,968$ ways of selecting 11 risky choices from 41. The computational time to compute Table 1 is around 8 hours for a sufficiently high degree of accuracy. To do this for all ${}_{41}C_{11} = 3,159,461,968$ possible selections is obviously an impossible task. Accordingly, selections were chosen at random and the necessary calculations performed for each selection. A subset of the results is presented in Table 2. The tremendous variability in the number of implied permissable combinations, in the extent of the overlap between theories and the number of implied non-dominating pairs should be noted. It should also be noted that selection number 18 in Table 2 is conspicuous for the large number of permissable combinations, by the relatively small amount of overlap and for the fact that the 11 basic risky choices implied 30 non-dominating pairwise choices.

The random selections shown in Table 2 is just a subset of the random selections that

Table 2: The implications of various random selections of 11 basic risky choices

selection number	Responses consistent with models						total number of possible responses	number of non-dominating pairs
	1	2	3	4	5	6		
1	162	44	26	5	1	14	252	19
2	64	12	11	1	7	7	102	11
3	77	18	19	10	2	9	135	15
4	84	42	22	13	2	12	175	18
5	18	2	3	3	0	5	31	5
6	87	23	10	6	1	9	136	15
7	61	13	9	0	5	7	95	10
8	97	25	8	4	6	9	149	13
9	70	51	38	9	1	14	183	21
10	43	12	8	2	5	6	76	10
11	103	28	12	5	4	9	161	13
12	81	29	14	5	0	11	140	14
13	94	23	19	8	8	11	163	17
14	61	12	4	6	0	5	88	8
15	104	42	14	9	8	12	189	16
16	56	30	10	4	7	10	117	13
17	56	23	8	1	1	5	94	11
18	253	106	30	22	2	16	429	30
19	158	45	17	4	1	13	238	19
20	66	19	7	3	0	7	102	12

were investigated. However, selection 18 (as it is numbered in Table 2) emerged as probably the best from all these selections. Accordingly it was used in the experiment: the 11 basic risky choices formed the 11 questions to be ranked in order in the Complete Ranking part of the experiment and the implied 30 non-dominating pairwise choices formed the Pairwise Choice part of the experiment. The actual 11 choices, as has already been pointed out, are the 11 choices indicated in Figure 2.

4.2 The Complete Ranking part of the Experiment

The discussion above has concerned the number of possible responses on the Pairwise Choice part of the experiment. However it clearly follows, since the basic 11 risky choices are common to both parts of the experiment, that exactly the same number of permissable responses for each theory are possible on the Complete Ranking part of the experiment.

There is no need to repeat the analysis given above. For each permissable response (under a particular theory) for the Pairwise Choice part, there corresponds a region of the relevant parameter space, and hence corresponds a permissable response on the Complete Ranking part of the experiment.

5 Implementing the Experiment

All the discussion above has concerned the choice of the risky choices in the experiment. The choice of the outcomes x_1, x_2 and x_3 now needs to be discussed as well as the details of the experiment. It was decided to employ a method previously used, and one that seems particularly suited to the problem under consideration, of choosing large payoffs but then selecting just one subject for playing out for real. Accordingly x_1, x_2 and x_3 were chosen to be £0, £200 and £1000. Advertisements were placed around the York campus and a total of 208 subjects were recruited. They took part in the Pairwise Choice part of the experiment at a time of their choosing and in the computer laboratory of **EXEC**, and were then given the instructions for the Complete Ranking part of the experiment, which they were to do at leisure and outside the laboratory. The instructions are reproduced in the Appendix. Subjects were instructed to report to a particular lecture theatre on 'Payday' and to bring their Complete Ranking with them at that time. They had to hand in their Complete Ranking as they entered. Of the original 208 subjects, a total of 179 attended 'Payday'. The analyis is restricted to these 179 subjects.

The payoff mechanism is as described in the instructions. For the Pairwise Choice part of the experiment, one subject was selected at random and then one of the 30 pairwise choice questions was selected at random. The previously-expressed choice of that subject on that question was then recalled and the preferred risky choice was then played out by the subject. Specifically, given that all risky choices were represented in the form of segmented circles, a hard copy version of the chosen risky choice was placed on a continuous roulette wheel and the chosen subject spun the wheel. The subject was paid the outcome. For the Complete Ranking part of the experiment, again one subject was chosen at random, and then two of the 11 risky choices were selected at random. The chosen subject's previously-

stated preferences were then consulted and the one of the two randomly-selected risky choices that was highest in the subject's complete ranking was played out - in the manner described above - and the subject paid the outcome.

6 Analysing the Results

The results are straightforward to present, particularly for the Complete Ranking part of the experiment: *no* subject had a ranking that was consistent with *any* of the theories. Accordingly, the value of Selten's r for all the theories is zero. The implied value of Selten's Measure of Predictive Success for the various theories is therefore just the negative of the value of Selten's a. These were reported above and it is clear that the largest value of $-a$ is equal to the smallest value of a which is that for Expected Utility theory. On the basis of Selten's Measure therefore, Expected Utility theory emerges as the best. One may, however, well be dissatisfied with this conclusion - particularly given the fact that no observation was consistent with any theory.

The situation with the Pairwise Choice part of the experiment was more encouraging: a total of 33 (out of the 179) subjects gave responses consistent with one or more of the preference functionals. Details are given in Table 3. The implications for Selten's Measure are given in Table 4. It is clear that Disappointment Aversion theory emerges as the best theory on the basis of this Measure.

7 Conclusions

The evidence for the usefulness of Selten's Measure is somewhat mixed. On the Complete Ranking part of the experiment, the Measure seems to have little value - because none of the observations are consistent with any of the theories. On the Pairwise Choice part, things are a little better - with Disappointment Aversion emerging as the front runner - it is parsimonious relative to the other generalisations of Expected Utility theory and it scores highly in terms of predictability. Setting aside Expected Utility theory, which is obviously more parsimonious than the other theories, Table 4 shows that Disappointment Aversion

Table 3: Choices consistent with the various theories in the Pairwise Choice part of the experiment

number of subjects	choices	**EU**	**DA**	**PR**	**RP**	**RQ**	**WU**
4	01010011111101000111111100001	*	*	*	*	*	*
1	01000011011101000111111100001	*	*	*	*	*	*
4	01010011111101000111111100001		*				
4	01010011111101000111111100001		*				
4	01010011111101000111111100001		*				
1	01100011011101000111111100001		*				
4	01010011111101000111111100001		*	*			
1	01100011011101000111111100001		*	*	*	*	
4	01010011111101000111111100001		*				
4	01010011111101000111111100001		*		*		
1	01000011011101000111111100001			*			
1	01100010010101100111111100001			*	*	*	

Table 4: Calculation of the Selten Measure for the Pairwise Choice Experiment

Model	No of allowable responses	No of observed responses	Selten's Measure
EU	16	5	0.02793
DA	67	31	0.17318
PR	161	12	0.06704
RP	156	11	0.06145
RQ	199	7	0.03911
WU	150	5	0.02793
totals	2^{30}	179	

is best both on the basis of r and of a - so, in practice, no trade-off, of the type provided by Selten's Measure, is needed on this occasion. Obviously this is not a finding that can be generalised. In comparing Expected Utility with Disappointment Aversion however, there is a trade-off, and Selten's Measure comes into its own. Nevertheless the feeling remains that with the very small values for r (the 'hit rate', to use Selten's expression), the Measure continues to suffer because it omits any consideration of observations 'close' to consistency with theories. It is also interesting to note that Disappointment Aversion theory, the front-runner in this analysis, consistently fares relatively badly when some form of maximised log-likelihood is used to compare the competing theories. But this, of course, is another story.

Finally, it should be noted that if the two parts of the experiment are combined - which is something that can, and perhaps should be, done because the same subjects performed both parts, then the conclusion must be reached that no subjects had behaviour in the experiment as a whole consistent with any of the theories. In this situation Selten's Measure tells us to choose the most parsimonious theory - namely Expected Utility theory. However, one might think that this conclusion is too brutal - when based on a Measure which considers all observations inconsistent with a theory equally 'bad'.

References

Bissey, M. E. (1997). Semi-parametric estimatiom of preference functionals. unpublished.

Camerer, C. F. (1995). Individual decision making. In Kagel, J. H. and Roth, A. E., editors, *Handbook of Experimental Economics*, pages 587–703. Princeton University Press.

Hey, J. D. (1997). Experiments and the economics of individual decision making. In Kreps, D. M. . and Wallis, K. F., editors, *Advances in Economics and Econometrics*, pages 171–205. Cambridge University Press.

Hey, J. D. (1998). An application of Selten's 'Measure of Predictive Success'. *Mathematical Social Sciences*, 35:1–16.

Karmarkar, U. (1978). Subjectively weighted utility: a descriptive extension of the expected utility model. *Organisational behavior and Human Performance*, 21:61–72.

Karmarkar, U. (1979). Subjectively weighted utility and the Allais paradox. *Organisational behavior and Human Performance*, 24:67–72.

Machina, M. (1982). 'Expected Utility' analysis without the independence axiom. *Econometrica*, 50:277–323.

Selten, R. (1991). Properties of a Measure of Predictive Success. *Mathematical Social Sciences*, 21:153–167.

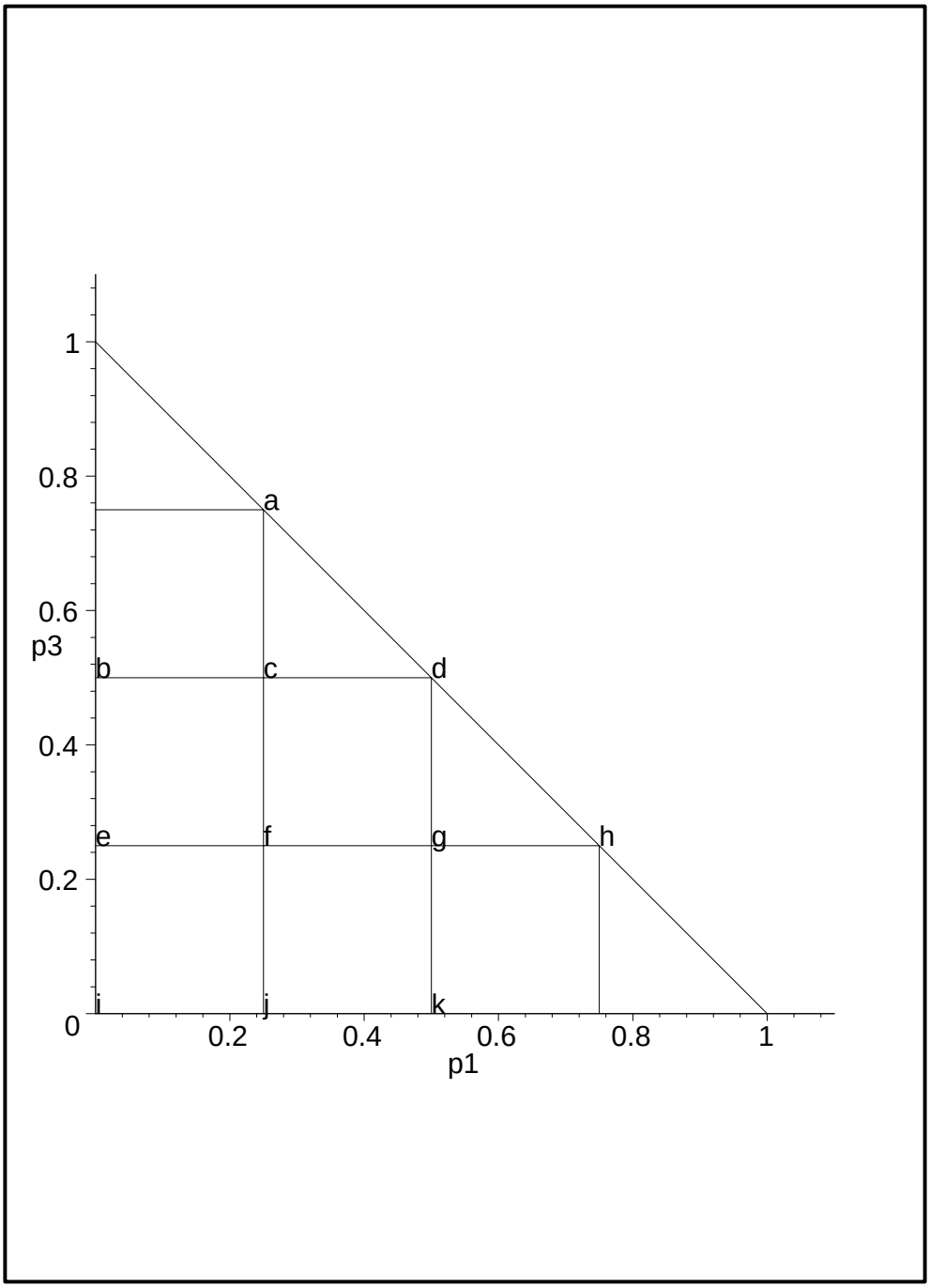


Figure 1: The risky choices in the earlier experiment

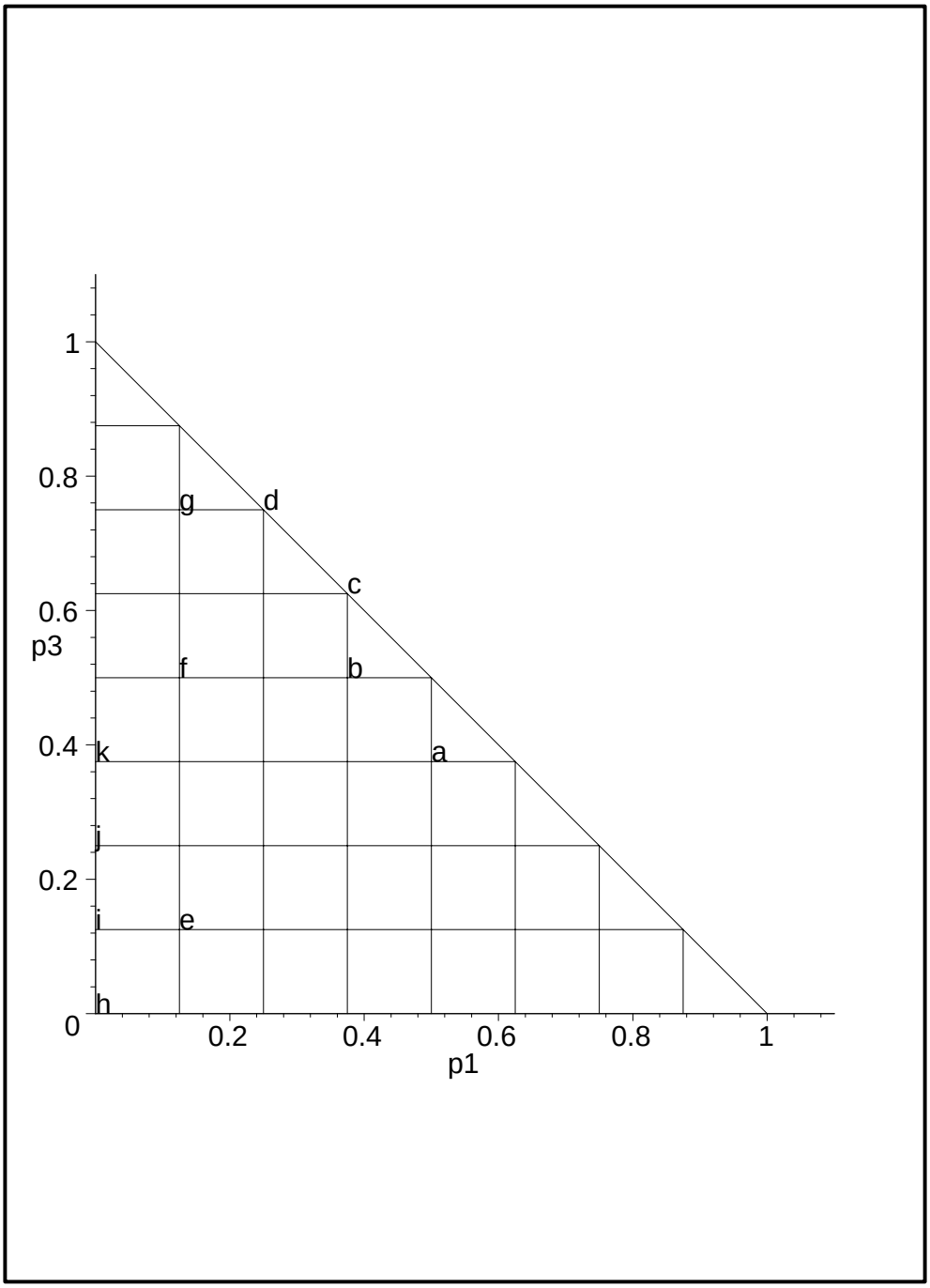


Figure 2: The risky choices in this experiment

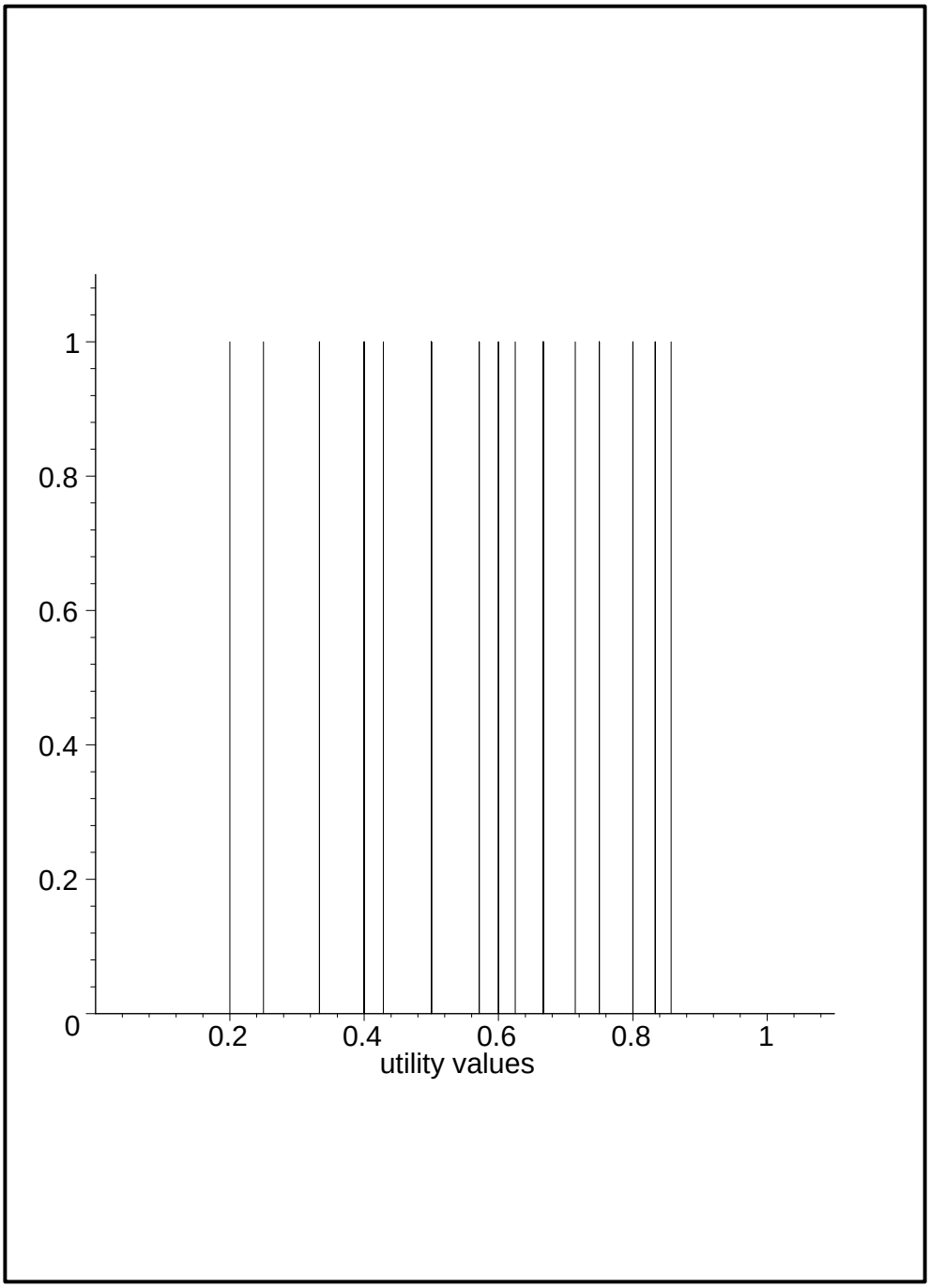


Figure 3: Expected Utility boundaries

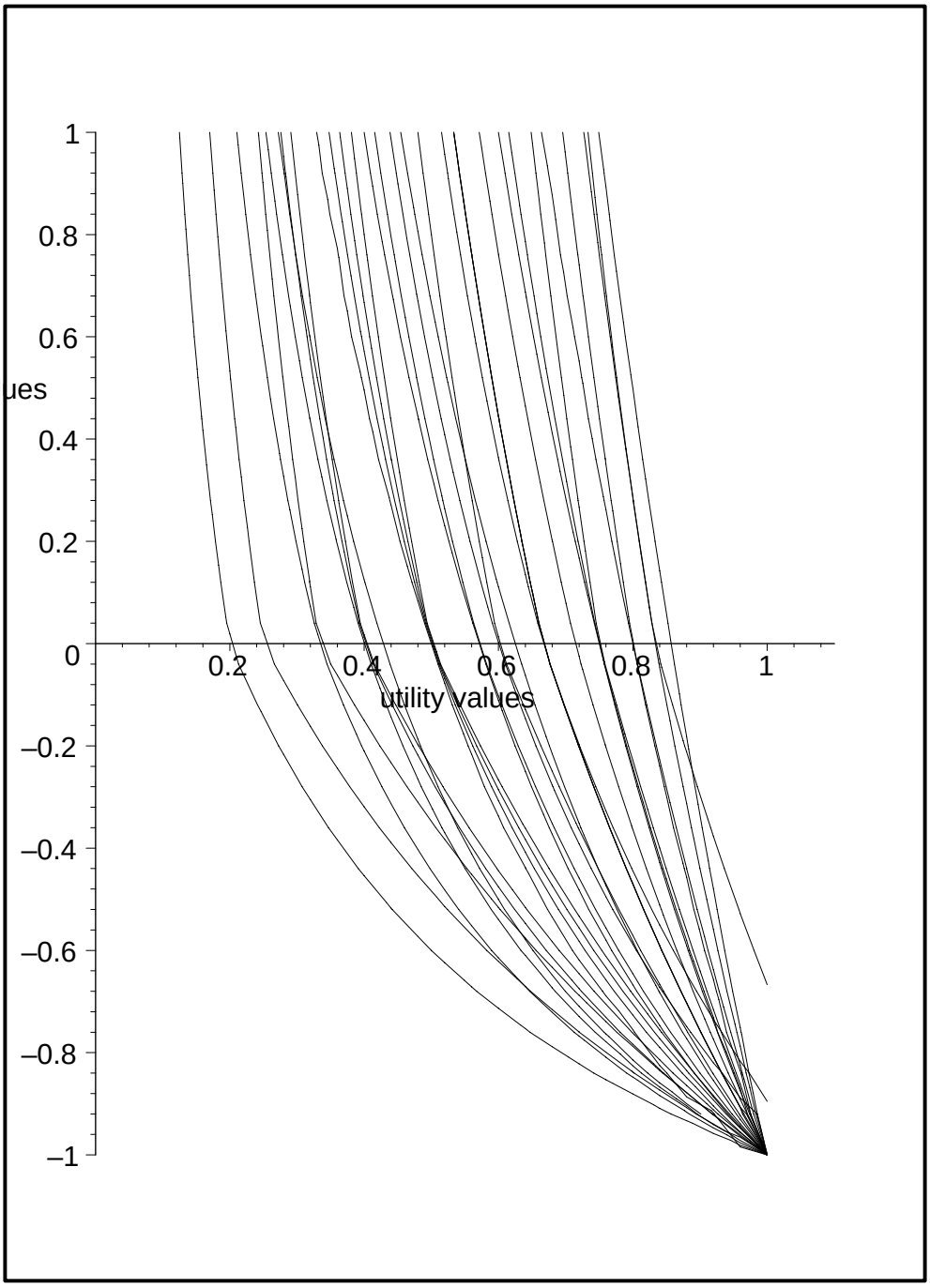


Figure 4: Disappointment Aversion boundaries

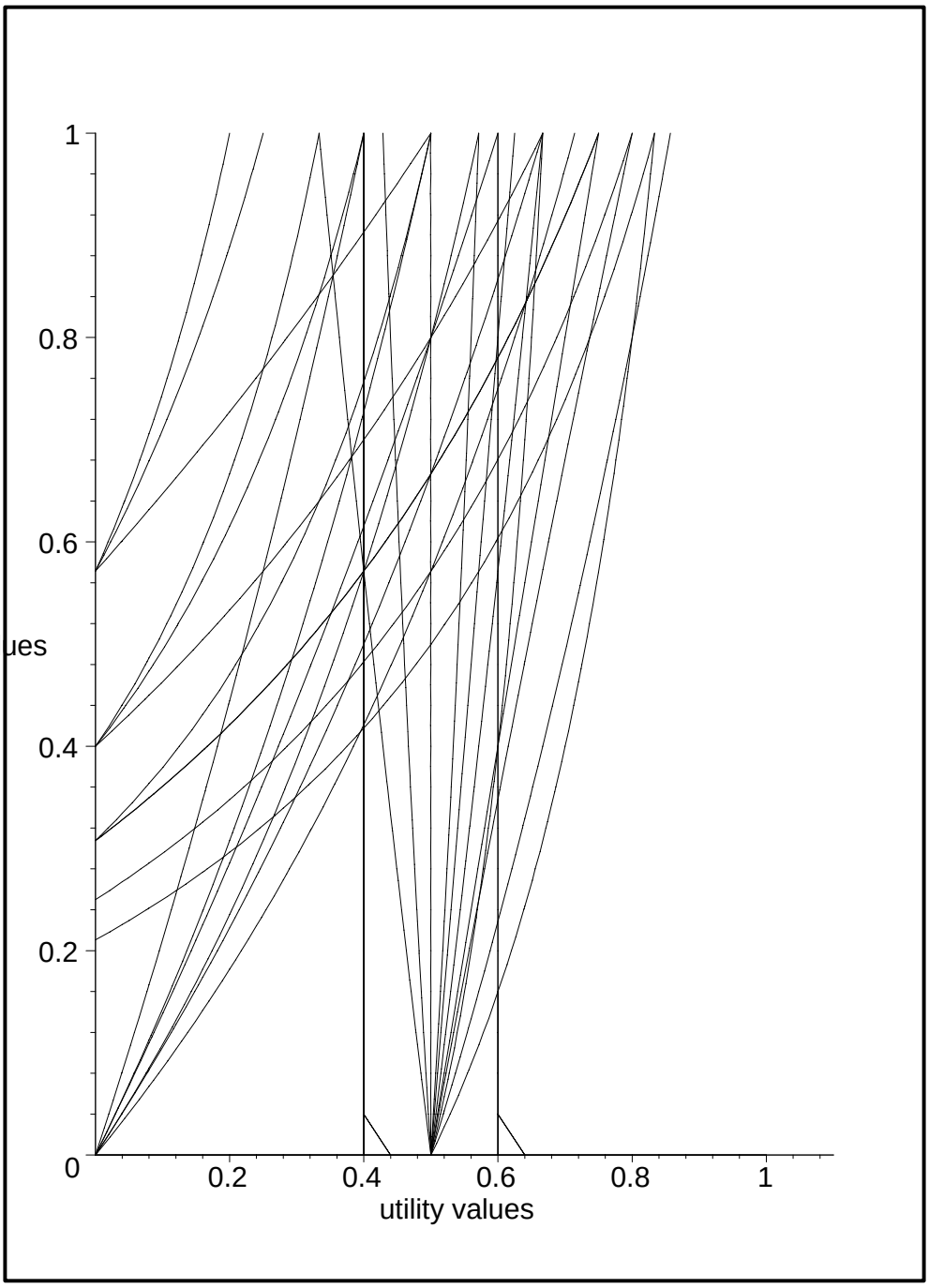


Figure 5: Prospective Reference boundaries

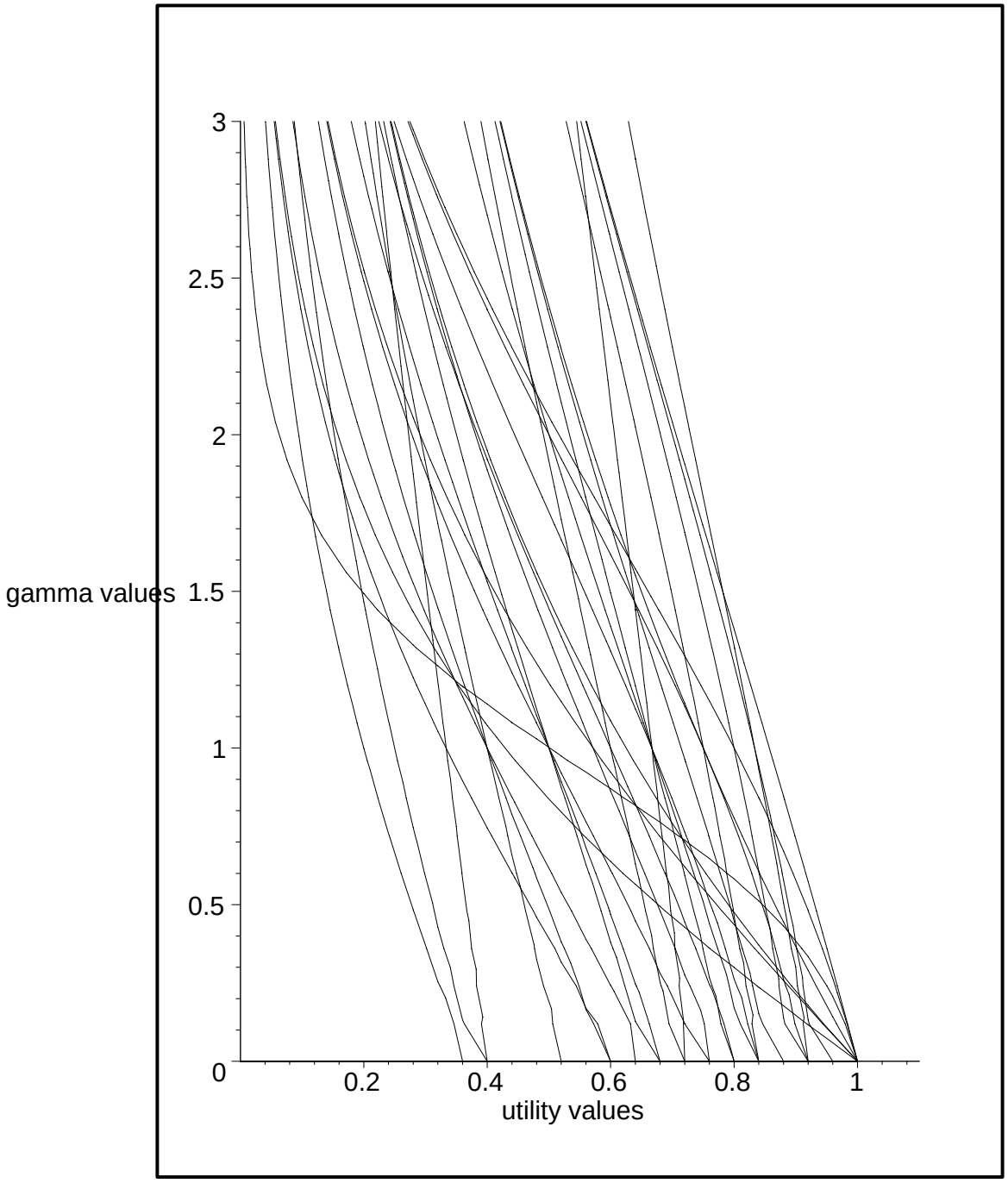


Figure 6: Rank Dependent (with Power weighting function) boundaries

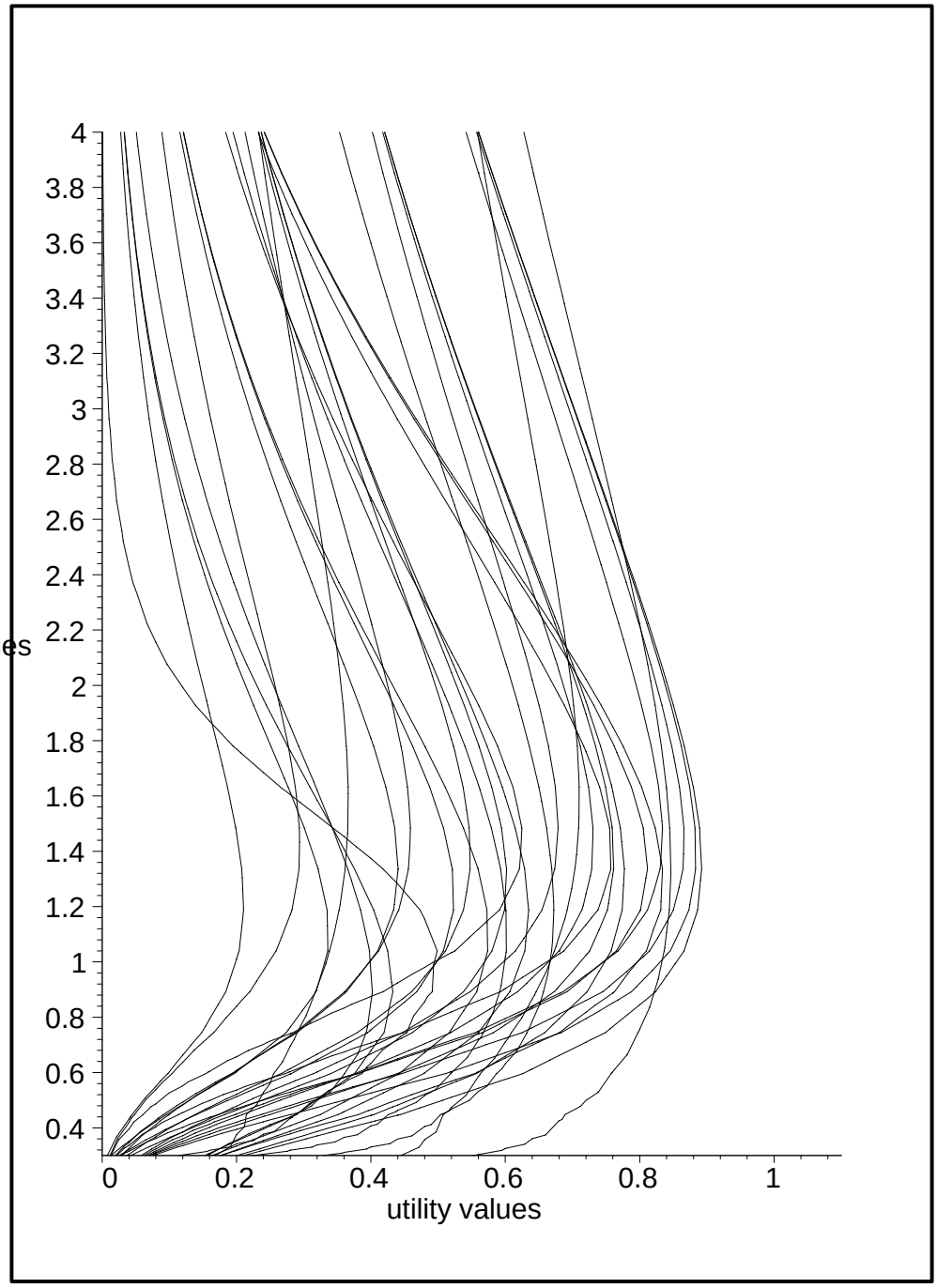


Figure 7: Rank Dependent (with Quiggin weighting function) boundaries

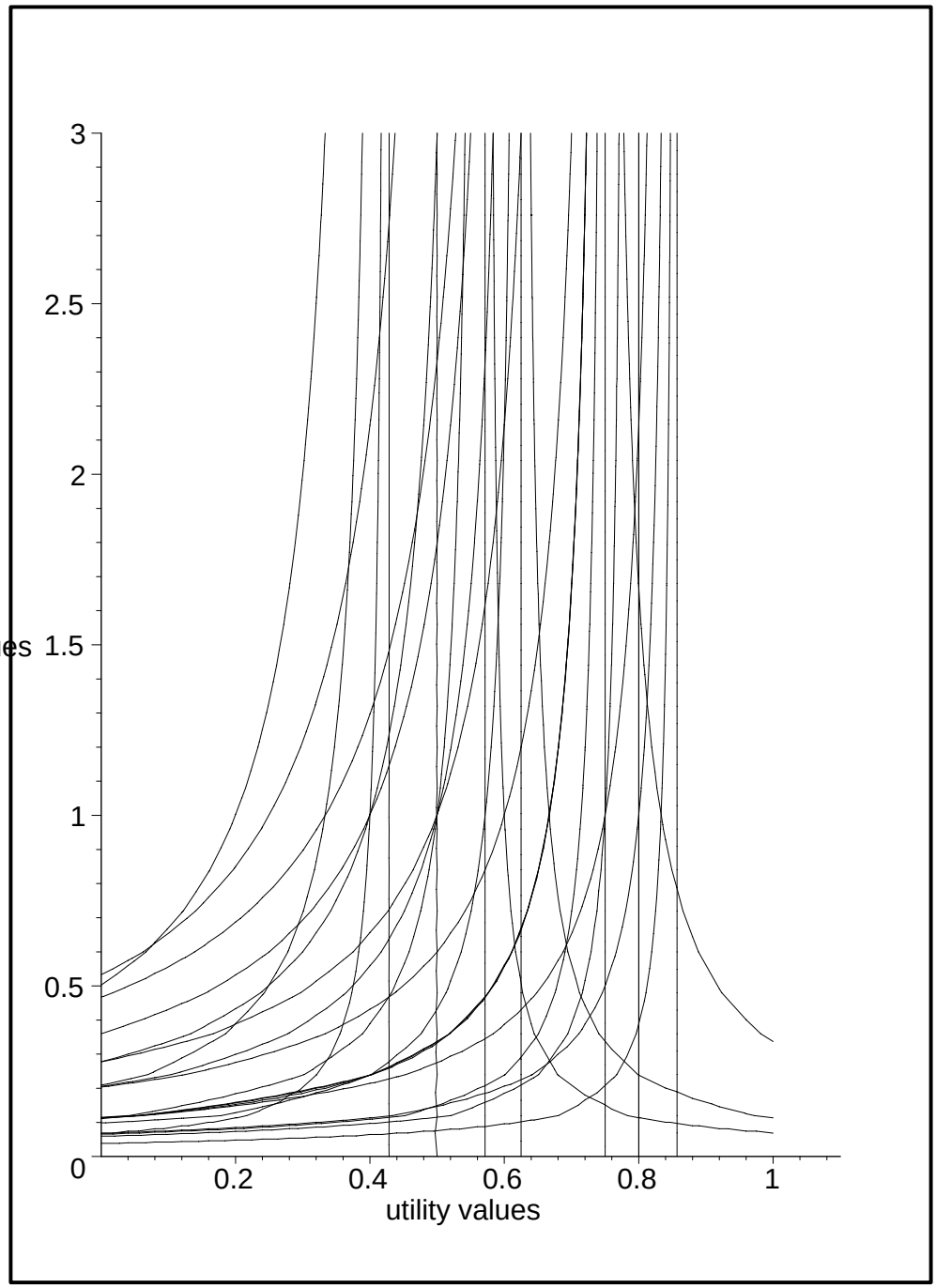


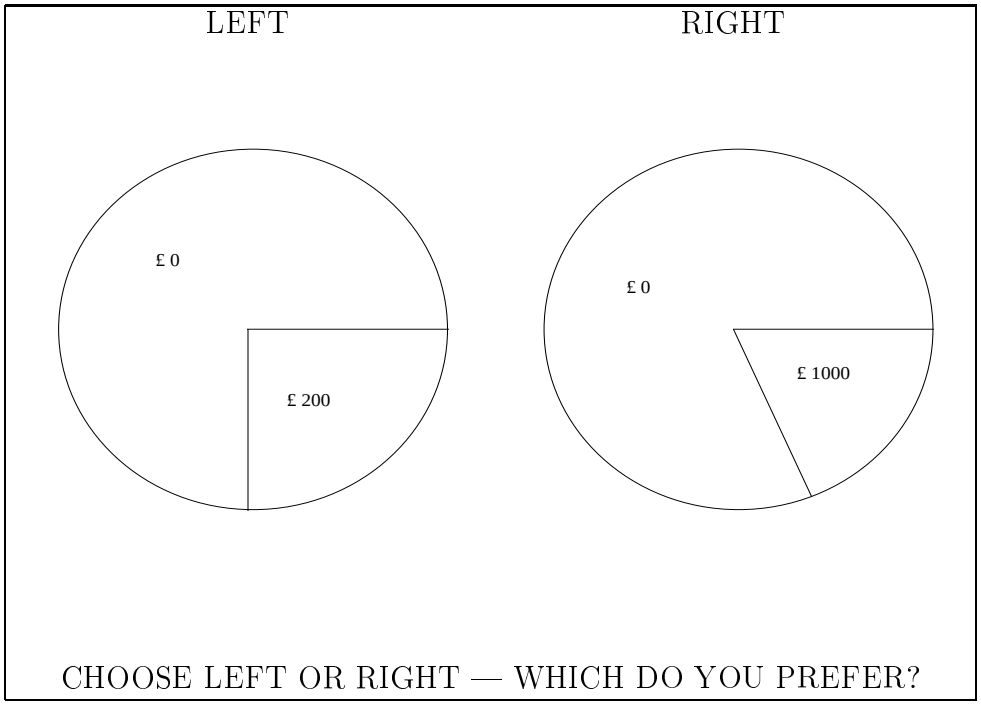
Figure 8: Weighted Utility boundaries

INSTRUCTIONS FOR EXPERIMENT 1/97

Experiment 1/97 consists of two parts, one computerised, the other not. Each participant must complete both parts before 'Payday'.

Part 1: choosing

This part is held in the EXEC laboratory (Derwent D Block, room D202). It consists in pairwise choice questions, each asking which of two risky prospects you personally prefer. In each question, the two prospects are represented in the form of circles, with particular segments representing particular outcomes. All outcomes in this experiment will be one of three amounts: £0, £200, £1000.
Here is how a question will appear on the computer screen:



The reward mechanism is described below. You have to complete the second part of the experiment before 'Payday'.

Part 2: ranking

At the end of part 1 of the experiment, a sheet of paper will be given to you with a set of risky prospects similar to the ones you have had to choose from in part 1 of the experiment. Each one of these prospects will be indexed by a letter. Once again, the possible outcomes will be £0, £200 and £1000.
What you have to do is the following: decide which one of these prospects you prefer the most, and report the corresponding letter. Then you have to decide which one you prefer next and report the corresponding letter; and so on until you report the prospect you prefer the least.
You will have to hand in the completed part 2 at the beginning of 'Payday'.

Payment mechanism

'Payday' will be held on Monday Week 6 (17th of November), at 5:15pm in room PX001 (Physics). You will be asked to hand in part 2 of the experiment as you come in. If you cannot attend 'Payday' in person, please send a representative (with your answers to part 2 of the experiment).
The reward mechanism is the following:

- For part 1 of the experiment, one of the participants present will be chosen at random. One of the questions will then be chosen also at random and the lottery this participant has reported as the most preferred will be played out for real. Accordingly, the participant will receive £0, £200 or £1000.
  In the event that the participant gets £0, the whole of the above procedure will be repeated (except that the previously chosen participant(s) will be excluded from future random draws), until someone has won either £200 or £1000.

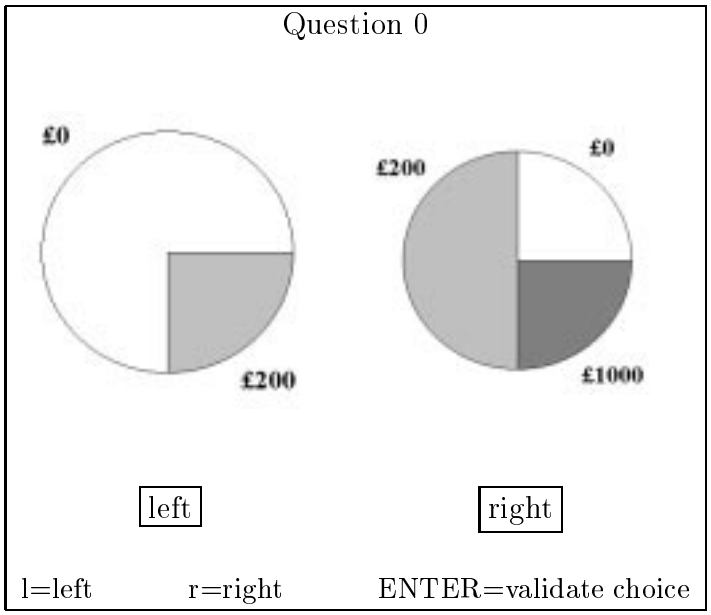
- For part 2 of the experiment, one of the participants will be chosen at random (participants having been picked up in part 1 being excluded from the draw); then two of the prospects he or she had to rank will be chosen at random. The prospect which is the highest in this participant's ranking is played out for real. As for part 1, the procedure will eventually be repeated until a participant gets £200 or £1000.

The payoff (either £200 or £1000) will be paid in cash immediately.

WELCOME TO PART 1 OF EXPERIMENT 1/97

Here is what is going to happen:

- When you face the welcome screen on the computer, press C to start.
- You are asked to write your name.
- The program takes you then through instructions for the experiment. Please read them carefully.
- If you have any questions at the end of the instructions, call the experimenter.
- The experiment consists of 30 questions of the following form:



- After having answered the 30 questions, please call the experimenter who will give you then part 2 of the experiment and the relevant instructions.

INSTRUCTIONS FOR EXPERIMENT 1/97: PART 2

At the end of part 1 of the experiment, a sheet of paper has been given to you with a set of 11 risky prospects similar to the ones you have had to choose from in part 1 of the experiment. Each one of these prospects is indexed by a letter (from A to K) above it. Once again, the possible outcomes are £0, £200 and £1000.
What you have to do is the following: decide which one of these prospects you prefer the most, and report the corresponding letter in the corresponding box at the bottom of the sheet. Then you have to decide which one you prefer next and report the corresponding letter; and so on until you report the prospect you prefer the least.
You will have to hand in the completed part 2 at the beginning of 'Payday'.

PAYMENT MECHANISM

'Payday' will be held on Monday Week 6 (17th of November), at 5:15pm in room PX001 (Physics). You will be asked to hand in part 2 of the experiment as you come in. If you cannot attend 'Payday' in person, please send a representative (with your answers to part 2 of the experiment).

> *Remember that if we do not have your answers to part 2 of the experiment, you cannot participate in the draw.*

The reward mechanism is organised as follows:

- For part 1 of the experiment, one of the participants present or represented will be chosen at random. This is done by selecting randomly one of the answer sheets handed in at the begining of 'Payday'.
  The chosen participant will then choose one question at random and the prospect this participant reported earlier as the most preferred will be played out for real.
  Playing out a prospect for real is done by placing a representation of the prospect on a roulette wheel which the participant spins. The segment of the prospect where the wheel stops decides the payoff.
  Accordingly, the participant will receive £0, £200 or £1000.
  In the event that the participant gets £0, the whole of the above procedure will be repeated (except that the previously chosen participant(s) will be excluded from future random draws), until someone has won either £200 or £1000.

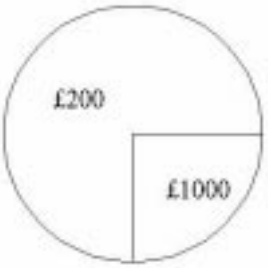
- For part 2 of the experiment, one of the participants will be chosen at random (participants having been picked up in part 1 being excluded from the draw); then two of the prospects he or she had to rank will be chosen at random. The prospect which is the highest in this participant's ranking will then be played out for real. As for part 1, the procedure will eventually be repeated until a participant gets £200 or £1000.

The payoff (either £200 or £1000) will be paid in cash immediately.

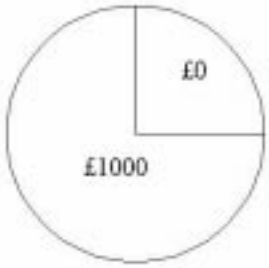
EXPERIMENT 1/97: PART 2 — COMPLETE RANKING
[The letter corresponding to each prospect is ABOVE it]

EMAIL	
NAME	
NB	

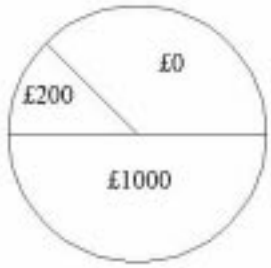
Prospect A



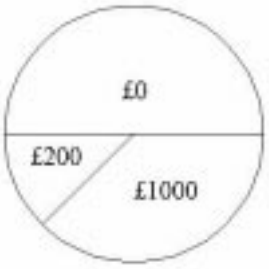
Prospect B



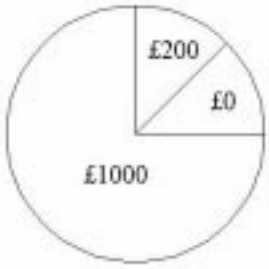
Prospect C



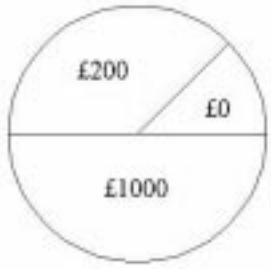
Prospect D



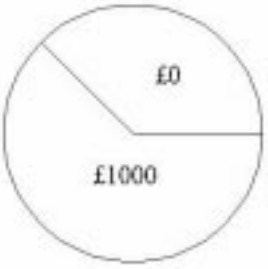
Prospect E



Prospect F

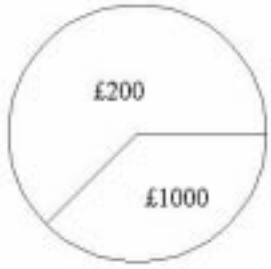


Prospect G

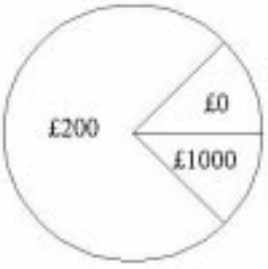


Prospect H

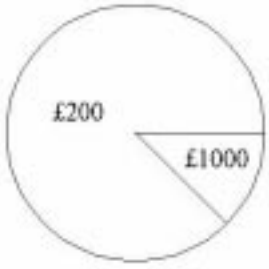
Prospect I



Prospect J



Prospect K



Please indicate your order of preference from the one you like the MOST to the one you like the LEAST

most	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	least